

Efficient Joint Network-Source Coding for Multiple Terminals with Side Information*

Chen Avin[†], Michael Borokhovich[†], Asaf Cohen[†] and Zvi Lotker[†]

November 28, 2011

Abstract

Consider the problem of source coding in networks with multiple receiving terminals, each having access to some kind of side information. In this case, standard coding techniques are either prohibitively complex to decode, or require network-source coding separation, resulting in sub-optimal transmission rates. To alleviate this problem, we offer a joint network-source coding scheme based on matrix sparsification at the code design phase, which allows the terminals to use an efficient decoding procedure (syndrome decoding using LDPC), despite the network coding throughout the network. Via a novel relation between matrix sparsification and rate-distortion theory, we give lower and upper bounds on the best achievable sparsification performance. These bounds allow us to analyze our scheme, and, in particular, show that in the limit where all receivers have comparable side information (in terms of conditional entropy), or, equivalently, have weak side information, a vanishing density can be achieved. As a result, efficient decoding is possible at all terminals simultaneously. Simulation results motivate the use of this scheme at non-limiting rates as well.

*The material in this paper was presented in part at the Information Theory and Applications Workshop, San Diego, USA, 2011 and the IEEE International Symposium on Information Theory, Saint Petersburg, Russia, 2011.

[†]Authors are with the Department of Communication Systems Engineering, Ben-Gurion University of the Negev, Israel. Email: {avin,borokhom,coasaf,zvilo}@bgu.ac.il

1 Introduction

In this work, we consider the problem of efficient distributed source coding for large networks with multiple receiving terminals. Assume that information source X is available at some source node s in a network with noiseless links. The source is to be distributed (with arbitrarily small error probability) to *several* receiving nodes in the network, $T = \{t_1, \dots, t_K\}$, where each node $t \in T$ has some side information Y^t available. We assume that for each t , (X, Y^t) have some known joint distribution. An example is given in Figure 1.

This problem arises in a multitude of networking applications, such as sensor networks, peer-to-peer, and content distribution networks. A few examples to be kept in mind can be a sensor in a network distributing a temperature measurement while each receiving node has its own altitude reading (which is highly correlated) or a video file streamed distributively in a network where receivers may have previous/noisy/low-resolution versions of that stream.

The problem of lossless coding with side information has been studied extensively. We give here only a brief overview of the closely related works. In [1], Slepian and Wolf considered the problem of separately encoding two correlated sources and joint decoding (Figure 2(a)). The asymmetric case where side information Y is available at the decoder is a special case (Figure 2(b)). In [2], Ho et al. considered the multicast problem with correlated sources (Figure 3(a)), and completely characterized the *rate region* for this problem: the set of required link capacities to support the multicast. In a way, [2] can be viewed as extending the Slepian Wolf problem to arbitrary networks through *network coding*. Further extensions also appeared in [3] and [4] (Figure 3(b)).

The canonical, three-node network of Slepian and Wolf [1], as well as its extension in [2] do not consider practical decoding algorithms. Such algorithms were considered in [5] using Turbo codes, in [6] based on the Wyner scheme [7] and in [8, 9, 10] based on Low-Density Parity-Check Codes (LDPC), including a solution for any point on the rate region [11]. In a recent study [12, 13], the authors use linear-programming based decoding techniques. An algebraic (Reed-Solomon-based) coding scheme was suggested in [14]. Nevertheless, a

major disadvantage of applying decoding techniques based on structured linear codes to a networked environment is the need to separate network and source coding. This separation can fail for most demand structures [15] (that is, will require higher rates than the necessary cut-set bounds). Hence, while the distributed source coding model was studied extensively in the information theory literature throughout the years, a huge gap still exists when trying to apply the *efficient decoding schemes to large networks with many terminals*, where network coding is essential to achieve maximal throughput.

First introduced by Ahlswede et al. in [16], network coding deals with various coding operations that can be performed at *intermediate nodes* in the network in order to achieve certain rate goals. For linear network coding [17, 18], decoding sums up to solving a set of linear equations. When all terminals can solve and reconstruct the original data, the application layer above the network code can choose the *source coding scheme* independently, as first the network code is decoded, and then the source code. This is a *separation based* coding scheme. However, in general, separation schemes fail to achieve the maximal throughput in the network (e.g., a rank equal to the max-flow at each node [17]), and an extra rate is required if each terminal wishes to decode its own subset of bits. In this case we say that *separation fails* [15]. The solution is thus one of the following: use extra rate (compared to the max-flow bound), or use *joint network-source code*, where the network and source codes are matched and are designed to be decoded together. To date, joint network-source coding schemes are prohibitively complex, while efficient decoding schemes require a structured source code, which is shattered by the lack of separation. A midpoint approach was taken in [19] using minimum-cost optimization, giving a trade-off between joint and separate network-source coding. In [20], a different approach was taken, yet still sub-optimal in terms of the required rates. Alternatively, in [21, 22] the authors suggest a sum-product-based algorithm, nevertheless, in this case, efficient decoding requires a network with bounded degree nodes. Hence, our goal in this work is to design efficient coding schemes which are both rate-optimal and applicable to general network topology.

Main Contribution: We formally define the setting for joint network-source coding in a network with side information at multiple terminals. We design a joint coding scheme that, under certain constraints, both achieves the lower bounds on the link capacities and facilitates efficient decoding at the terminals, by inducing, at all terminals simultaneously, a source code which is based on a low-density parity-check matrix, and hence can be decoded efficiently. This joint network-source code is achieved by a sparsification procedure of the source coding matrix and the network transfer matrix together, at the code-design stage. Our analysis of the sparsification performance is based on a novel connection between matrix sparsification and rate-distortion theory, a connection which besides being interesting on its own, allows us to both analyze the best possible sparsification performance and give a randomized algorithm to approximate it. Numerical results also illustrate the efficiency of the derived codes and algorithms. Furthermore, this connection gives a novel randomized algorithm for coding with a fidelity criterion.

The rest of the paper is organized as follows. Section 2 gives the required notation, related results and describes the model of distributed source coding with several receivers, including a discussion of the main difficulties. Section 3 gives our main results. Section 4 describes the coding algorithm. Section 5 gives the numerical results and Section 6 concludes the paper.

2 Preliminaries

Network and Source Model: A network is defined as a directed acyclic graph $(\mathcal{V}, \mathcal{E})$, where \mathcal{V} is the set of vertices (nodes) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges (links). Associated with each edge $e \in \mathcal{E}$ is a capacity $c(e) \geq 0$. For any node v , we denote the set of incoming and outgoing edges of node v by $In(v)$ and $Out(v)$, respectively.

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed random variables with alphabet \mathcal{X} . Source $\{X_i\}$ is available at node $s \in \mathcal{V}$. The case of multiple sources can be dealt with similar techniques, by coding for points on the rate region. Let $T \subseteq \mathcal{V}$,

$|T| = K$, be a set of terminals. A terminal $t \in T$ has (side) information $\{Y_i^t\}_{i=1}^\infty$ available to it. For each t , $\{Y_i^t\}_{i=1}^\infty$ is a sequence of independent and identically distributed random variables with alphabet \mathcal{Y} . We assume that the pairs (X_i, Y_i^t) may be dependent, with some known joint distribution $p(x, y^t)$, yet for different time indices i and j , (X_i, Y_i^t) and (X_j, Y_j^t) are independent (that is, a memoryless model). We assume that the node s has no incoming edges and that each terminal node in T has no outgoing edges.

For any vector of rates $(R_e)_{e \in \mathcal{E}}$, a $((2^{nR_e})_{e \in \mathcal{E}}, n)$ code comprises the following mappings: $g_e^n : \mathcal{X}^n \mapsto \{1, \dots, 2^{nR_e}\}$, for $e \in \text{Out}(s)$ and $g_e^n : \Pi_{e' \in \text{In}(v)} \{1, \dots, 2^{nR_{e'}}\} \mapsto \{1, \dots, 2^{nR_e}\}$, for $e \in \text{Out}(v), v \neq s$. That is, the source codes a block of n inputs and maps it to messages on its outgoing links. Each internal node may *code* its inputs and, again, map them to messages on the outgoing links. Thus, for each terminal $t \in T$, $\Pi_{e \in \text{In}(t)} \{1, \dots, 2^{nR_e}\}$ is the set of inputs to that terminal. For simplicity, we denote this input by $f_t(X^n)$. $f_t(X^n)$, together with $(Y_1^t, \dots, Y_n^t) \triangleq Y^{t,n}$ is the data the terminal uses for decoding.

Throughout this work, we use *linear network coding*, that is, source symbols are mapped to a vector of length w over some finite field F . An edge $e \in \text{Out}(v), v \in \mathcal{V}$, carries an element of the field, which is a linear combination (over F) of the variables on all edges $e' \in \text{In}(v)$. Note that this “unit capacity” assumption does allow for any rational capacity by adding parallel edges and normalizing over several network uses. It is convenient to represent the input to the source as w imaginary incoming edges. Under this notation, f_e denotes the global coding vector for edge e (its coefficients with respect to the input variables) and for some $t \in T$, V_t denotes the linear space spanned by $\{f_e : e \in \text{In}(t)\}$.

For nodes s and t , we denote by $\text{maxflow}(s, t)$ the capacity of the maximal flow between s and t . This flow is equal to the capacity of the minimal cut separating s from t .

Definition 1 ([23]). A w -dimensional linear network code on an acyclic network is *linear broadcast* if for every non-source node t , $\text{rank}(V_t) = \min\{w, \text{maxflow}(s, t)\}$.

Corollary 1 ([23]). *For any acyclic network and sufficiently large base field F , there exists a w -dimensional linear broadcast code.*

In words, for any dimension w , there exist a large enough base field F and a linear network code such that for any non-source node t , we have $f_t(\mathbf{s}) = \mathbf{s}M_t$, where \mathbf{s} is the w dimensional source vector over F and M_t is a $w \times |In(t)|$ matrix over F , whose columns are the global coding vectors $\{f_e : e \in In(t)\}$. Furthermore, $\text{rank}(M_t) = \min\{w, \text{maxflow}(s, t)\}$.

As the base field F should be large enough to allow for the matrices $\{M_t\}$ to satisfy Corollary 1, one cannot assume that the source alphabet, \mathcal{X} is of the same size as F . From this point on, we assume that $\mathcal{X} = \{0, 1\}$ and that $F = \text{GF}(2^m)$ for sufficiently large m . In this case, wm consecutive input bits are mapped to a source vector \mathbf{s} .

Proposition 1. *Let $\mathcal{N} = (\mathcal{V}, \mathcal{E}, s, T)$ be a network with a source $s \in \mathcal{V}$ and terminals $T \subseteq \mathcal{V} \setminus s$. Assume $(\mathcal{V}, \mathcal{E})$ is a directed acyclic graph and that $c(e) = 1$ for any $e \in \mathcal{E}$. Set $w = \max_{t \in T} \text{maxflow}(s, t)$ and a sufficiently large m . Then assuming bits b_1, \dots, b_{mw} are available at the source s , there exists a linear network code over base field $\text{GF}(2^m)$ such that any terminal $t \in T$ receives $m \cdot \text{maxflow}(s, t)$ linearly independent equations on b_1, \dots, b_{mw} .*

In other words, if bits b_1, \dots, b_{mw} are available at the source, a terminal t with $\text{maxflow}(s, t) = l \leq w$, can receive (possibly after linear operations at that terminal) a vector of ml bits, $\mathbf{y} = (\tilde{b}_1, \dots, \tilde{b}_{ml})$ which is the result of $(b_1, \dots, b_{mw})B$, for $B \in \{0, 1\}^{wm \times lm}$ with rank ml . The proof of Proposition 1 is rather technical and appears in Appendix A.

Linear Codes for Structured Binning: Consider the source coding problem in Figure 2(a). By [1], the set of achievable rates is given by $R_1 \geq H(X|Y)$, $R_2 \geq H(Y|X)$ and $R_1 + R_2 \geq H(X, Y)$. In the particular case where the side information Y is available at the decoder (Figure 2(b)), a rate $H(X|Y)$ is required. Now, assume X is a binary symmetric random variable, related to Y through a binary symmetric channel. That is, $Y = X \oplus E$, where E is a binary random variable with $\Pr(E = 1) = p$ and \oplus denotes XOR operation. Note that in this model $H(X|Y) = h(p)$, where h is the binary entropy function $h(p) = -p \log p - (1 - p) \log(1 - p)$. In this case, a coding scheme based on linear codes was suggested by Wyner [7] (see [24] also).

Fact 1. *Let $\epsilon > 0$ be arbitrary. For n sufficiently large, there exists an $n \times m$ matrix H with $m < (h(p) + \epsilon)n$ and a function $f : \{0, 1\}^m \mapsto \{0, 1\}^n$ such that $\Pr(f(e^n H) \neq e^n) \leq \epsilon$.*

The source vector x^n is thus encoded as $x^n H$, while, at the decoder, $e^n H$ is calculated according to $e^n H = x^n H \oplus y^n H$ (as y^n is available) and x^n is then reconstructed using $x^n = y^n \oplus f(He^n)$. Hence, the decoding process is analogous to the decoding of a linear code with a parity check matrix H . For example, in [8], LDPCs are used. As a result, H is also required to be sparse and of a certain structure.

Distributed Source Coding with Multiple Terminals: We now turn to our original problem. In terms of the achievable rates alone (regardless of the possibility for efficient decoding) the problem can be viewed as multicast in the presence of side information. The following can be seen as a corollary of [4].

Corollary 2. *Let $\mathcal{N} = (\mathcal{V}, \mathcal{E}, s, T)$ be a network with a source $s \in \mathcal{V}$ and terminals $T \subseteq \mathcal{V} \setminus s$. Assume $(\mathcal{V}, \mathcal{E})$ is a directed acyclic graph. Let source X be available at s and for each t , side information Y^t be available at $t \in T$. Then, a necessary and sufficient condition for X to be reconstructed at the terminals with an arbitrarily small probability of error is $\text{maxflow}(s, t) \geq H(X|Y^t)$ for all t .*

A possible scheme for achieving the bound in Corollary 2 is *random binning* at the source with *linear network coding* [4]. In this case, it is easy to construct a source coding scheme which is oblivious to the network code, in the sense that a terminal t does not care whether it receives the k most significant bits of a bin index, or any k linearly independent equations on the $n > k$ bits of the bin index. Another possible scheme uses linear coding and minimum entropy decoding [2]. Note, however, that the decoding complexity of both schemes is exponential in the block length. Should, however, one try to use linear codes with a structure that facilitates efficient decoding, for more than two terminals the structure of the code will be destroyed by the network code.

To see this, assume bits x^n are available at the source. Denote by w_t the max-flow to terminal t and by w the maximal max-flow among all terminals, $\max_t w_t$. By Corollary 2, there exists a (joint) network-source coding scheme for which each terminal t can decode x^n with an arbitrarily small (as $n \rightarrow \infty$) error probability as long as $w_t \geq H(X|Y^t) + \epsilon$. However, due to the lack of network-source coding separation [15], even if source node s creates K nested codewords, were the codeword intended for t is of length $nH(X|Y^t)$ (assuming $nH(X|Y^t)$ is an integer) and is a subset of the codewords for terminals with higher rates, in general, there does not exist a network code to ensure that each terminal t indeed receives the subset of bits *intended to it* while satisfying the min-cut bounds (this can be done, in general, for at most two terminals [25]). By Proposition 1, however, there exists a network code for which terminal t' does receive $nH(X|Y^{t'})$ *linearly independent equations* on the $n \max_t H(X|Y^t)$ transmitted bits. In other words, source node s can generate a single codeword of length $n \max_t H(X|Y^t)$, using a good parity check matrix H of size $n \times n \max_t H(X|Y^t)$. Denote this codeword by $\mathbf{c} = x^n H$. \mathbf{c} is hence of length $n \max_t H(X|Y^t)$. Each m bits in \mathbf{c} are mapped to a symbol in $\text{GF}(2^m)$, and a vector of w symbols, \mathbf{s} , is transmitted through the network. At terminal t , the received vector is $\mathbf{y} = \mathbf{s} M_t$. We assume $w m$ divides $n \max_t H(X|Y^t)$, so \mathbf{c} might be transmitted using several network uses. By Proposition 1, the $nH(X|Y^t)$ bits received at t can be represented as $(b_1, \dots, b_{nH(X|Y^t)}) = x^n H B_t$, where B_t is an $n \max_{t'} H(X|Y^{t'}) \times nH(X|Y^t)$ matrix with full rank (if $|In(t)| > w_t$, one can take w_t linearly independent coding vectors and discard the rest).

While H is a parity-check matrix designed to facilitate efficient decoding of e^n from $e^n H$, B_t is a matrix *defined by the network code*, and may be different for each terminal t . As a result, even if H , for example, is sparse, $H B_t$ might not be sparse at all and lack any structure that allows efficient decoding. This is the reason we say that the structure of the source code is shattered by the network code.

3 Efficient Decoding Using LDPC

The focus is thus on designing H (the source code), a linear network code with matrices B_t , $1 \leq t \leq K$, and possibly post-processing matrices P_t , such that, for each terminal t , $\bar{H}_t = HB_tP_t$ is a good parity-check matrix for terminal t , with a structure that facilitates efficient decoding. In this section, we center our attention on H and the P_t 's, and derive sufficient conditions to guarantee that each terminal t sees, with high probability, a low-density parity-check matrix \bar{H}_t , assuming B_t is the result of either deterministic or random linear network coding.

At the heart of the suggested scheme is a matrix sparsification algorithm and its analysis for high rates (almost square matrices). In short, for each HB_t , we wish to find a matrix P_t such that HB_tP_t is sparse. In [26, 27], the authors consider randomized matrix sparsification algorithms. If A is an $n \times (n - k)$ matrix, the authors seek an $(n - k) \times (n - k)$ matrix P such that AP is sparse. The analysis therein, however, aims at bounding the difference between the sparsity achieved by the algorithm and the best possible performance, without quantifying the best possible performance directly. In the next subsection, we show a novel connection between matrix sparsification and rate distortion theory. Through this connection, we are able to give bounds on the best possible performance directly, and, as a result, give sufficient conditions under which one can indeed find such P which yields a sparse matrix AP . At the basis of our results are Lemmas 1 and 2, which give lower and upper bounds on the sparsification performance. Theorem 1 below, our main result in this section, utilizes these lemmas to show that indeed, as long as the difference in the strength of the side information (that is, conditional entropies) available at the nodes is small, designing joint network-source codes which induce low-density parity-check matrices at the terminals is possible.

Theorem 1. *Let $(\mathcal{V}, \mathcal{E}, s, T)$ be a network with binary uniform source X available at node s and terminals $t \in T$ with side information Y^t . Assume $\text{maxflow}(s, t) \geq R_t = H(X|Y^t) + \epsilon$. Then, for large enough block length n , it is possible to construct a joint source and network*

code such that each terminal can decode X using a syndrome resulting from a parity check matrix \bar{H}_t with normalized density at most $D(\max_{t'} R_{t'} - R_t) + O(\frac{\log n}{n})$, where $D(\cdot)$ is the distortion rate function of a binary uniform random variable under Hamming distortion.

This result gives a vanishing density whenever the strength of the side information at the nodes is approximately the same. In particular, at the limit of high rates, a vanishing density can be achieved. A small (but still useful for efficient decoding) density can be achieved when the variation in the rates is slightly larger. Moreover, note that a naive matrix sparsification based on Gauss elimination would result only in a linear rate-sparsity trade-off, as the lower part of the matrix (of size $n(1 - R) \times nR$) is arbitrary. For $R \rightarrow 1$, the improvement over Gauss elimination is arbitrarily large. To see this, note that the derivative of $D(R)$ tends to 0 as $R \rightarrow 1$, compared to the $1/2$ rate achieved by Gauss elimination. For example, taking $R = 1 - O(\frac{(\log n)^2}{n})$ results in a density of $O(\frac{\log n}{n})$. All this can be clearly seen from Figure 4, which compares $D(R)$ to Gauss elimination, together with the results of the sparsification algorithm we used.

Theorem 1 results from analyzing both the properties of the joint network-source code HB_t seen by terminal t , and the best possible sparsification performance. The first part is rather technical and is summarized in Proposition 2 at the end of this section. We now focus on the second part and show that, at least in the randomized setting, performance guarantees for matrix sparsification can indeed be achieved via the distortion-rate function.

3.1 Analyzing Matrix Sparsification Via Rate-Distortion

Matrix sparsification is formally defined as follows.

Problem 1. Given a full rank matrix $A \in \text{GF}(q)^{n \times (n-k)}$, find an invertible matrix $P \in \text{GF}(q)^{(n-k) \times (n-k)}$ which minimizes the number of non-zeros (nnz) in AP .

In [26, 27] the authors prove that matrix sparsification is NP-hard, and give approximation algorithms. Note that, in our model, matrix sparsification is part of the *code design*, and

not the decoding process. Hence, with successful sparsification, efficient decoding is possible throughout the transmission of the (possible many) source blocks.

The analysis in [26, 27] bounds the performance compared to the optimum, but does not include any guarantee on the achieved sparsity. In the context of our problem, such a guarantee is essential in order to verify that each terminal actually receives a syndrome which can be seen as created by a low-density matrix.

First, a few definitions are required. For any $a, b \in \text{GF}(q)$ define $d(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b. \end{cases}$

With a slight abuse of notation, for any $\mathbf{a}, \mathbf{b} \in \text{GF}(q)^n$ define $d(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^n d(\mathbf{a}(i), \mathbf{b}(i))$. It is common to denote $d(\mathbf{a}, \mathbf{b})$ as $\text{nnz}(\mathbf{a} - \mathbf{b})$. At the heart of the matrix sparsification algorithm stands the following problem.

Problem 2. *Min-Unsatisfy:* For any $A \in \text{GF}(q)^{n \times (n-k)}$ and $\mathbf{b} \in \text{GF}(q)^n$, find the vector $\mathbf{x} \in \text{GF}(q)^{n-k}$ which minimizes $d(A\mathbf{x}, \mathbf{b})$.

Let A_{-j} denote the matrix A with the j -th column, \mathbf{a}_j , removed. Under these definitions, a possible matrix sparsification algorithm has the following form [26].

Algorithm 1. For each $1 \leq j \leq m$:

1. Let $\mathbf{x} = \text{Min-Unsatisfy}(A_{-j}, \mathbf{a}_j)$.
2. Replace the column \mathbf{a}_j with $\mathbf{a}_j - A_{-j}\mathbf{x}$.

The following lemma bounds the performance of *any algorithm* for solving Problem 1.

Lemma 1. *Let $A \in \text{GF}(q)^{n \times (n-k)}$ be a random matrix with i.i.d. entries. Let $P(A) \in \text{GF}(q)^{(n-k) \times (n-k)}$ be any invertible matrix, whose entries may depend on those of A . The expected number of non-zeros in AP satisfies $\frac{1}{(n-k)n} \mathbb{E} \{ \text{nnz}(AP) \} \geq D((n-k)/n)$, where $D(\cdot)$ is the distortion-rate function of $A(1, 1)$ under Hamming distortion.*

Proof. The proof is based on the converse to the rate-distortion theorem (e.g., [28, Section 10.4]). In particular, we show that for any possible choice of P (which, in general, depends on the *realization* of A), each column of AP is at most as sparse (in expectation) as the

vector of differences between a column of A and the closest vector to it among the *best* rate-distortion code of size q^{n-k-1} .

Let P be a given invertible matrix over $\text{GF}(q)$. For each column \mathbf{p}_j in P , denote by $i^*(j)$ the index of the first non-zero entry in \mathbf{p}_j and let $\lambda_j = 1/P(i^*(j), j)$. Let Λ denote the diagonal matrix whose diagonal entries are $\{\lambda_j\}_{j=1}^{n-k}$. Finally, denote the minimizer in Problem 2 by $\hat{\mathbf{x}}_A(\mathbf{b})$, that is, $\hat{\mathbf{x}}_A(\mathbf{b}) = \arg \min_{\mathbf{x}} d(A\mathbf{x}, \mathbf{b})$. We have:

$$\begin{aligned} \frac{1}{(n-k)n} \mathbb{E}\{\text{nnz}(AP)\} &= \frac{1}{(n-k)n} \mathbb{E}\{\text{nnz}(AP\Lambda)\} \\ &= \frac{1}{(n-k)n} \mathbb{E}\left\{\sum_{j=1}^{n-k} \text{nnz}(A\mathbf{p}_j\lambda_j)\right\} = \frac{1}{(n-k)n} \sum_{j=1}^{n-k} \mathbb{E}\left\{d\left(\mathbf{a}_{i^*(j)}, -\sum_{l=i^*(j)+1}^{n-k} \mathbf{a}_l P(l, j)\lambda_j\right)\right\} \\ &\geq \frac{1}{(n-k)n} \sum_{j=1}^{n-k} \mathbb{E}\left\{d\left(\mathbf{a}_{i^*(j)}, -A_{-j}\hat{\mathbf{x}}_{A_{-j}}(\mathbf{a}_{i^*(j)})\right)\right\}. \quad (1) \end{aligned}$$

Now, simply consider $\mathbf{a}_{i^*(j)}$ as a source vector of length n , and the set $\{A_{-j}\mathbf{x} : \mathbf{x} \in \text{GF}(q)^{n-k-1}\}$ as a rate-distortion code of rate $(n-k-1)/n$ (that is, having q^{n-k-1} codewords if A_{-j} has full rank). The structure of the code is, of course, defined by the matrix A_{-j} . By the converse to the rate-distortion theorem, for *any* subset $\{\mathbf{b}_1, \dots, \mathbf{b}_{q^{nR}}\} \subseteq \text{GF}(q)^n$, where $0 \leq R \leq 1$, we have

$$\mathbb{E}\{d(\mathbf{a}_{i^*(j)}, \mathbf{b}_i)\} \geq nD(R), \quad (2)$$

where \hat{i} is the minimizer of $d(\mathbf{a}_{i^*(j)}, \mathbf{b}_i)$ over all $1 \leq i \leq q^{nR}$. A-fortiori, (2) holds if the set $\{\mathbf{b}_1, \dots, \mathbf{b}_{q^{nR}}\}$ is not chosen optimally (herein, it is chosen as the linear subspace spanned by the columns of A_{-j}). As a result,

$$\begin{aligned} \frac{1}{(n-k)n} \sum_{j=1}^{n-k} \mathbb{E}\left\{d\left(\mathbf{a}_{i^*(j)}, -A_{-j}\hat{\mathbf{x}}_{A_{-j}}(\mathbf{a}_{i^*(j)})\right)\right\} &\geq \frac{1}{(n-k)n} \sum_{j=1}^{n-k} nD\left(\frac{n-k-1}{n}\right) \\ &\geq D\left(\frac{n-k}{n}\right), \end{aligned}$$

where the last inequality follows from the monotonicity of $D(R)$. \square

Note that the above proof is based on the fact that a solution to Problem 1 can be seen as sequentially applying Problem 2 to each of the columns of A , with the rest of the columns as the second argument. This is also the reason that an approximation for Problem 2 results in an approximation for Problem 1, and bounds on the possible performance in Problem 2 give bounds on the performance in Problem 1.

In fact, in certain cases, the bound in Lemma 1 is achievable. For binary random matrices with uniform i.i.d. entries, we have the following achievability result.

Lemma 2. *Let $A \in GF(2)^{n \times (n-k)}$ be a random matrix with uniform i.i.d. entries. For any $0 \leq D \leq 1/2$ which satisfies $\frac{n-k}{n} \geq 1 - h(D) + 2\frac{\log n}{n} + \frac{1}{n}$, there exists an invertible matrix $P \in GF(2)^{(n-k) \times (n-k)}$ such that with probability at least $1 - (n-k)2^{-n}$, for all $1 \leq j \leq n-k$, we have $\text{nnz}(A\mathbf{p}_j) \leq nD$. In particular, $\frac{1}{n(n-k)}\text{nnz}(AP) \leq D$.*

Proof. At the basis of the proof stands the fact that for binary uniform random variables, linear codes achieve the rate-distortion function [29].

First, we show how to construct a square matrix P which achieves the required sparsity of AP . For any matrix $C \in GF(2)^{n \times nR}$, $0 \leq R \leq 1$, define $\mathcal{C} = \{C\mathbf{x} : \mathbf{x} \in GF(2)^{nR}\}$. In [29], it is shown that if C is random with uniform i.i.d. entries, and \mathbf{b} is a binary vector with uniform i.i.d. entries, then if $R \geq 1 - h(D) + 2\frac{\log n}{n}$ we have

$$\Pr \left(\min_{\mathbf{c} \in \mathcal{C}} d(\mathbf{c}, \mathbf{b}) > nD \right) \leq 2^{-n}. \quad (3)$$

Note that the codeword \mathbf{c} which minimizes $\min_{\mathbf{c} \in \mathcal{C}} d(\mathbf{c}, \mathbf{b})$ is simply a linear combination of the columns of C , that is, $C\hat{\mathbf{x}}_C(\mathbf{b})$. Hence, to construct the invertible matrix P , we proceed inductively as follows. We consider \mathbf{a}_1 as the source vector (\mathbf{b} in (3)) and A_{-1} as the linear code. Note that both are random with uniform i.i.d. entries. Define a matrix S_1 such that all its columns except the first one are equal to those of the identity matrix. For the first column, set $S_1(1, 1) = 1$ and the rest of the values as simply $\hat{\mathbf{x}}_{A_{-1}}(\mathbf{a}_1)$. S_1 is invertible. To continue, construct S_j as follows. Set all its columns except the j -th equal to those of the

identity matrix. For the j -th column, set $S_j(j, j) = 1$ and the rest of the values as simply $\hat{\mathbf{x}}_{(AS_1 \cdots S_{j-1})_{-j}}((AS_1 \cdots S_{j-1})_j)$.

We set the matrix P as simply $S_1 \cdots S_{n-k}$. Clearly, P is invertible. Note that the j -th column of P is simply the j -th column of $S_1 \cdots S_j$, that is $\mathbf{p}_j = (S_1 \cdots S_j)_j$. Moreover, this column is chosen to minimize the sum of $(AS_1 \cdots S_{j-1})_j$ and the linear combination of the columns of $(AS_1 \cdots S_{j-1})_{-j}$.

We now show that for each j , $AS_1 \cdots S_{j-1}$ remains random with uniform i.i.d. entries. Each matrix S_i can be represented as a multiplication of $s_i - 1$ matrices, s_i being the number of non-zero entries in the i th column of S_i . Denote them by $W_{S_i}^1, \dots, W_{S_i}^{s_i-1}$. Each of these matrices differs from the identity matrix by one entry only: $W_{S_i}^l$ has 1 at the entry corresponding to the l th non-zero element of the i th column of S_i (excluding the diagonal element). For example, $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$. Thus, $S_1 \cdots S_{j-1} = W_{S_1}^1, \dots, W_{S_1}^{s_1-1} \cdots W_{S_{j-1}}^1, \dots, W_{S_{j-1}}^{s_{j-1}-1}$. Consider the multiplication $AW_{S_1}^1$. This operation simply replaces one column in A with its XOR with another column. Since for two independent uniform bits X and Y , X and $X \oplus Y$ are also independent uniform bits, $AW_{S_1}^1$ remains a matrix with binary uniform i.i.d. entries. By induction, $AS_1 \cdots S_{j-1}$ is also uniform with i.i.d. entries.

Utilizing the above, to compute the probability that some column of AP has a density higher than nD , we have

$$\begin{aligned} \Pr\left(\cup_{1 \leq j \leq n-k} \text{nnz}(A\mathbf{p}_j) > nD\right) &\leq \sum_{1 \leq j \leq n-k} \Pr\left(\text{nnz}(A\mathbf{p}_j) > nD\right) \\ &= \sum_{1 \leq j \leq n-k} \Pr\left(\text{nnz}(A(S_1 \cdots S_j)_j) > nD\right) \\ &= \sum_{1 \leq j \leq n-k} \Pr\left(\text{nnz}\left((AS_1 \cdots S_{j-1})_{-j} \cdot \hat{\mathbf{x}}_{(AS_1 \cdots S_{j-1})_{-j}}((AS_1 \cdots S_{j-1})_j)\right) > nD\right) \\ &= (n-k) \Pr\left(\text{nnz}(A_{-1} \hat{\mathbf{x}}_{A_{-1}}(\mathbf{a}_1)) > nD\right) \leq (n-k)2^{-n}, \quad (4) \end{aligned}$$

where the last inequality applies if $\frac{n-k-1}{n} \geq 1 - h(D) + 2\frac{\log n}{n}$. \square

Remark 1. Note that if each column in P is chosen *separately* of the others, in a way that the $n - k - 1$ entries of \mathbf{p}_j minimize the distance between \mathbf{a}_j and some linear combination of the columns in A_{-j} , then P is not necessarily invertible. However, it is not hard to see that all its off-diagonal entries are uniform i.i.d., while all the diagonal entries are deterministic and equal to 1. To compute the probability that P is invertible, note that the first row of P has 2^{n-k-1} possibilities. The i -th row of P cannot be a linear combination of the first $i - 1$ rows (there are 2^{i-1} such combinations), yet, since $P(i, i) = 1$, half of the combinations are not counted. As a result, the probability that such a P is invertible is

$$\frac{(2^{n-k-1})(2^{n-k-1} - 2^1/2)(2^{n-k-1} - 2^2/2) \dots (2^{n-k-1} - 2^{n-k-1}/2)}{2^{(n-k)^2 - (n-k)}} = \prod_{i=1}^{n-k-1} (1 - 2^{-i}) \approx 0.29.$$

3.2 Proof of Theorem 1

At the basis of the proof is the following theme: we identify the properties of the transfer matrix HB_t seen by a terminal t . As B_t is defined by the network code, hence constrained by the network topology and, in general, cannot be optimized to our needs, we seek a proper H such that, with high probability, either HB_t is sparse or it can be sparsified. Since the constraints under which HB_t may be sparse are not necessarily met (depending on B_t), we take the freedom of choosing H such that at least the result in Lemma 2 can be used, and HB_t can be sparsified.

The proof is based on the following proposition, which gives a qualitative description of the properties of the transfer matrices HB_t seen by each terminal t . The results herein are stronger than those required to use Lemma 2, but we include them for completeness. A proof is given in Appendix B.

Proposition 2. *Assume H is an $n \times (n - k)$ random binary matrix with i.i.d. components, where $n - k = \Theta(n)$ and $\Pr(H(1, 1) = 1) = \lambda/n$ for some $\lambda \geq 1$. Then,*

1. *If B_t is full rank then each column of HB_t is composed of independent bits (that is, column-wise independence).*

2. If B_t is full rank and H is chosen at random with uniform entries ($\lambda = n/2$), then HB_t has uniform i.i.d. entries.
3. If each column in B_t contains a linear (in n) number of non-zero entries, e.g., $c_j n$ then $\lim_{n \rightarrow \infty} \Pr((HB_t)(i, j) = 0) = \frac{1}{2} (1 + e^{-2c_j \lambda})$, where $(HB_t)(i, j)$ denotes the (i, j) th entry of HB_t . If, in addition, $\lambda = w(1)$, we have $\lim_{n \rightarrow \infty} \Pr((HB_t)(i, j) = 0) = \frac{1}{2}$.
4. If each column in B_t contains a sub-linear number of non-zero entries then $\lim_{n \rightarrow \infty} \Pr((HB_t)(i, j) = 1) = 0$.
5. Let \mathbf{b}_j and $\mathbf{b}_{j'}$ be two columns of B_t . If $d(\mathbf{b}_j, \mathbf{b}_{j'}) = \Theta(n)$, then by setting $\lambda = \omega(n)$ the random variables $(HB_t)(i, j)$ and $(HB_t)(i, j')$ are independent for each i . Hence, HB_t has independent columns. If, however, $d(\mathbf{b}_j, \mathbf{b}_{j'}) = o(n)$, then by choosing λ as some positive constant, we have $\lim_{n \rightarrow \infty} \Pr((HB_t)(i, j) \neq (HB_t)(i, j')) \rightarrow 0$.

Proof of Theorem 1. First, by Proposition 2, for each terminal t , the transfer matrix HB_t can either be made sparse using a proper choice of H , or it can be made completely i.i.d. with binary uniform entries. This is since $\lambda(n)$ is a parameter of the source code, and can be chosen by the code designer.

Then, at the “worst case” of HB_t being a uniform i.i.d. matrix, we use Algorithm 1, which results, at the limit of infinitely many repetitions, in a sparsity ratio of $2D(R)$ (Lemma 2 and [26]). Note that by [27, Theorem 8], there are unsparsifiable matrices, that is, matrices which cannot be made sparser than a linear rate-sparsity curve. In a sense, our results show that under a proper choice of H , we can guarantee that the transfer matrices seen by the terminals *can* be sparsified below the linear curve - up to the distortion-rate curve.

Remember that for $q = 2$ and $\mathbf{b}(0)$ being Bernoulli(p), we have

$$R(D) = \begin{cases} h(p) - h(D) & \text{if } 0 \leq D \leq \min\{p, 1 - p\} \\ 0 & \text{if } D > \min\{p, 1 - p\} \end{cases}$$

and $D(R)$ is the inverse of $R(D)$. Thus, for example, from analyzing $D(R)$ for $R \rightarrow 1$, we

have that for $R = 1 - \frac{(\log n)^2}{n}$, $D(R) = \frac{\log n}{n}$. □

4 A Randomized Algorithm for Linear Rate-Distortion

Note that through the relation between matrix sparsification and rate-distortion, we are now able to give a randomized algorithm to approximate the closest codeword in a rate-distortion problem (using linear codes) and assess its approximation factor compared to the optimum. Denote by $C(I, :)$ the matrix consisting of rows I of the matrix C , for some index set I . Consider the following algorithm, whose performance guarantee is a direct application of Lemma 2 and [26].

Algorithm 2. Input: A linear code \mathcal{C} of rate R (defined by an $n \times nR$ matrix C). A vector \mathbf{b} . Output: A code word \mathbf{c} .

1. Randomly choose an index set I of nR independent rows in C ,
2. Solve for \mathbf{x} in $C(I, :)\mathbf{x} = \mathbf{b}(I)$,
3. Return $C\mathbf{x}$.

Corollary 3. Assume C and \mathbf{b} are i.i.d. binary symmetric. With probability at least e^{-d} , Algorithm 2 returns a codeword \mathbf{c} whose distance from \mathbf{b} is at most $(\frac{nR}{d} + 2) \cdot nD(R)$.

5 Numerical Results

Figure 4 gives a numerical example of how close one can get to the rate-distortion curve using Algorithm 2. This algorithm is at the basis of the sparsification process and, in fact, the density of the matrices we use for Figure 5 is that depicted by the approximation curves in Figure 4. The sparsification algorithm used was Algorithm 1, with a random guess to solve the Min-Unsatisfy problem (similar to Algorithm 2).

To depict the performance of LDPC codes created from matrix sparsification of a random i.i.d. matrix with uniform bits, Figure 5 gives the bit error rates (BER) for decoding of such

codes. Figure 5 shows also a bit error rate for two matrices created by the following *structured* approach: an identity matrix is replicated 4×5 times, followed by about 10^5 permutations, where every permutation preserves the constant number of ones in each column (4 ones) and in each row (5 ones). While the results of LDPC codes created from matrix sparsification are inferior to well-optimized LDPCs, note that the codes suggested in this work do not require network-source coding separation, and hence, unlike optimized codes, can be used in networks with multiple receivers and without the (possibly very large) excess capacity above the min-cuts required when separation-based scheme is used. Sparse matrices can also be achieved for deterministic network coding matrices, resulting from the polynomial time algorithm of [30]. For example, sparsification results for a graph of 70 nodes, with coefficients over $\text{GF}(2^4)$, can be found in Figure 6. Note that the given percentage of non-zero elements is over $\text{GF}(2^4)$, and not over $\text{GF}(2)$. Figure 7 includes additional simulation results, for decoding of random LDPC's achieved by matrix sparsification algorithm - Algorithm 1. We can clearly see that better sparsification yields better bit error rate. Since for the same number of repetitions, Algorithm 1 gives better sparsification for smaller matrices (for larger matrix, more repetitions are needed to achieve better solution for the min-unsatisfy problem), the best BER is achieved by the smallest (128×115) matrix.

Figure 8 shows the density (of ones) achieved by the Algorithm 1, compared to Gauss elimination and the rate-distortion function $D(R)$, which is the lower bound (as we have shown in Lemma 1) for any matrix sparsification algorithm. We can see that a single repetition of min-unsatisfy in the Algorithm 1 results in almost the same sparsification achieved by the Gauss elimination. As the number of repetition grows, the density becomes much closer to the lower bound $D(R)$. The matrices used in the simulation are: 300×240 , and 300×270 which correspond to the rates 0.8 and 0.9 respectively.

On the more practical side, we note that further sparsification can be sought using methods to sparsify sparse matrices [31]. Moreover, the benefit in sparse network coding matrices goes beyond the problem of coding with side information. For example, more efficient algo-

rithms for solving linear equations (in the simple multicast scenario) can be used [32, 33].

6 Conclusion

In this work, we formally defined joint network-source coding for networks where multiple receivers have side information. We described a code-design procedure which is based on matrix sparsification, enabling each terminal in the network, at the limit of high rates, to receive codewords corresponding to low-density parity-check codes, thus facilitating efficient decoding. Since our scheme is not based on network-source coding separation, optimal rates, matching the cut-set bounds, can be achieved. Simulations performed depict encouraging results, also at non-limiting rates.

References

- [1] D. Slepian and J. Wolf, “Noiseless coding of correlated information sources,” *IEEE Trans. Inform. Theory*, vol. 19, no. 4, pp. 471–480, July 1973.
- [2] T. Ho, M. Médard, M. Effros, and R. Koetter, “Network coding for correlated sources,” in *CISS*, 2004.
- [3] J. Barros and S. D. Servetto, “Network information flow with correlated sources,” *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 155–170, January 2006.
- [4] M. Bakshi and M. Effros, “On achievable rates for multicast in the presence of side information,” in *Proc. ISIT*, Toronto, Canada, July 2008, pp. 1661–1665.
- [5] A. Aaron and B. Girod, “Compression with side information using turbo codes,” in *Data Compression Conference, Proceedings. DCC*. IEEE, 2002, pp. 252–261.
- [6] S.S. Pradhan and K. Ramchandran, “Distributed source coding using syndromes (DISCUS): Design and construction,” *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, 2003.
- [7] A. Wyner, “Recent results in the Shannon theory,” *Information Theory, IEEE Transactions on*, vol. 20, no. 1, pp. 2–10, 1974.
- [8] A.D. Liveris, Z. Xiong, and C.N. Georgiades, “Compression of binary sources with side information at the decoder using LDPC codes,” *Communications Letters, IEEE*, vol. 6, no. 10, pp. 440–442, 2002.

- [9] G. Caire, S. Shamai, and S. Verdú, “Lossless data compression with error correcting codes,” in *Information Theory, 2003. Proceedings. IEEE International Symposium on*, 2003, p. 22.
- [10] C.F. Lan, A.D. Liveris, K. Narayanan, Z. Xiong, and C. Georghiades, “Slepian-Wolf coding of multiple M-ary sources using LDPC codes,” in *Data Compression Conference, Proceedings. DCC*. IEEE, 2004, p. 549.
- [11] D. Schonberg, K. Ramchandran, and SS Pradhan, “Distributed code constructions for the entire Slepian-Wolf rate region for arbitrarily correlated sources,” in *Proceedings of Data Compression Conference, DCC*, 2004, pp. 292–301.
- [12] T.P. Coleman, A.H. Lee, M. Médard, and M. Effros, “On some new approaches to practical Slepian-Wolf compression inspired by channel coding,” in *Data Compression Conference, Proceedings. DCC*, 2004, pp. 282–291.
- [13] T.P. Coleman, M. Medard, and M. Effros, “Towards practical minimum-entropy universal decoding,” in *Data Compression Conference, Proceedings. DCC*, 2005, pp. 33–42.
- [14] S. Li and A. Ramamoorthy, “Algebraic codes for Slepian-Wolf code design,” in *Proc. of IEEE International Symposium on Information Theory*, Saint Petersburg, Russia, August 2011, pp. 1966–1970.
- [15] A. Ramamoorthy, K. Jain, P. A. Chou, and M. Effros, “Separating distributed source coding from network coding,” *IEEE Trans. Inform. Theory*, vol. 52, pp. 2785–2795, June 2006.
- [16] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, “Network information flow,” *IEEE Trans. Inform. Theory*, vol. 46, no. 4, pp. 1204–1216, July 2000.
- [17] S.-Y. R. Li, R. W. Yeung, and N. Cai, “Linear network coding,” *IEEE Trans. Inform. Theory*, vol. 49, no. 2, pp. 371–381, February 2003.
- [18] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, “A random linear network coding approach to multicast,” *IEEE Trans. Inform. Theory*, vol. 52, no. 10, pp. 4413–4430, October 2006.
- [19] A. Lee, M. Médard, K.Z. Haigh, S. Gowan, and P. Rubel, “Minimum-cost subgraphs for joint distributed source and network coding,” in *Proc. of Workshop on Network Coding, Theory and Applications*. Citeseer, 2007.
- [20] Y. Wu, V. Stanković, Z. Xiong, and S.-Y. Kung, “On Practical Design for Joint Distributed Source and Network Coding,” *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1709–1720, 2009.
- [21] G. Maierbacher, J. Barros, and M. Médard, “Practical source-network decoding,” in *Wireless Communication Systems, ISWCS 2009. 6th International Symposium on*. IEEE, 2009, pp. 283–287.

- [22] S. B. Cruz, G. Maierbacher, and J. Barros, “Joint Source-Network Coding for Large-Scale Sensor Networks,” in *Proc. of IEEE International Symposium on Information Theory*, Saint Petersburg, Russia, August 2011, pp. 484–488.
- [23] R.W. Yeung, *Information theory and network coding*, Springer Verlag, 2008.
- [24] R. Zamir, S. Shamai, and U. Erez, “Nested linear/lattice codes for structured multiterminal binning,” *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1250–1276, 2002.
- [25] E. Erez and M. Feder, “Capacity region and network codes for two receivers multicast with private and common data,” in *Workshop on Coding, Cryptography and Combinatorics*, 2003.
- [26] T. Neylon, *Sparse solutions for linear prediction problems*, Ph.D. thesis, New York University, 2006.
- [27] L.A. Gottlieb and T. Neylon, “Matrix sparsification and the sparse null space problem,” *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pp. 205–218, 2010.
- [28] T.M. Cover and J. Thomas, *Elements of information theory*, Wiley, 2006.
- [29] G. Cohen, “A nonconstructive upper bound on covering radius,” *Information Theory, IEEE Transactions on*, vol. 29, no. 3, pp. 352–353, May 1983.
- [30] S. Jaggi, P. Sanders, PA Chou, M. Effros, S. Egner, K. Jain, and L. Tolhuizen, “Polynomial time algorithms for multicast network code construction,” *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 1973–1982, 2005.
- [31] S.F. Chang and S.T. McCormick, “A hierarchical algorithm for making sparse matrices sparser,” *Mathematical Programming*, vol. 56, no. 1, pp. 1–30, 1992.
- [32] D. H. Wiedemann, “Solving sparse linear equations over finite fields,” *IEEE Transactions on Information Theory*, vol. 32, no. 1, pp. 54–62, January 1986.
- [33] Y.B. Gol’dshstein, “Portrait of the inverse of a sparse matrix,” *Cybernetics and Systems Analysis*, vol. 28, no. 4, pp. 514–519, 1992.

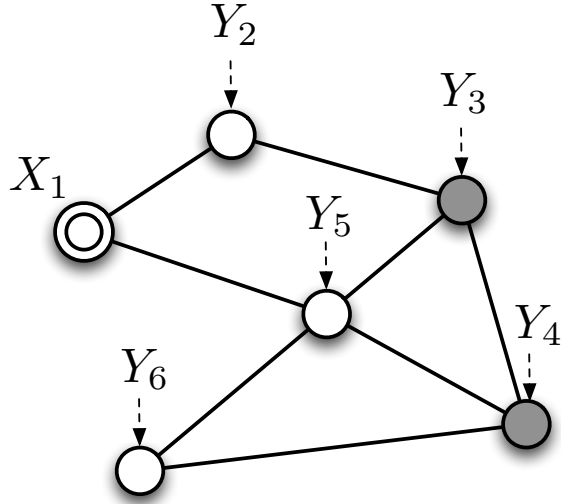
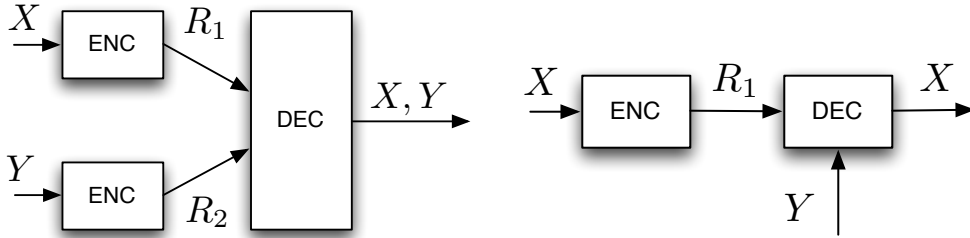


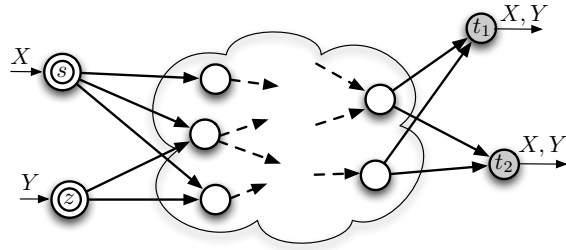
Figure 1: A network with side information. X_1 (the source) is demanded at Terminals 3 and 4. Each node has a side information message Y_i , correlated with X_1 .



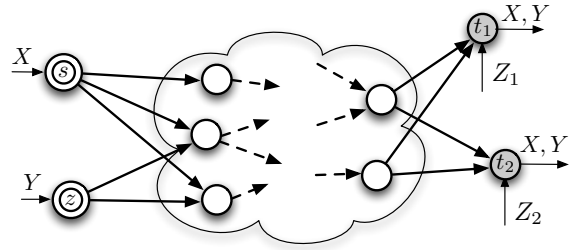
(a) The Slepian-Wolf network [1].

(b) The asymmetric Slepian-Wolf network.

Figure 2: (a) Upper encoder has the source X , lower one has the side information Y , which is, in general, correlated with X . We are interested in the set of rates (R_1, R_2) such that both X and Y can be reconstructed at the decoder. (b) The encoder describes the source X to a decoder which has side information Y available. We are interested in the rate R_1 such that X can be reconstructed at the decoder.



(a) A network with correlated sources [2].



(b) A network with correlated sources and side information [4].

Figure 3: (a) Nodes s and z have sources X and Y available. In general, X and Y are correlated. The sources are demanded (loselessly) at the terminals t_1 and t_2 . (b) In this case, terminals t_1 and t_2 have additional side information available to them, Z_1 and Z_2 , respectively.

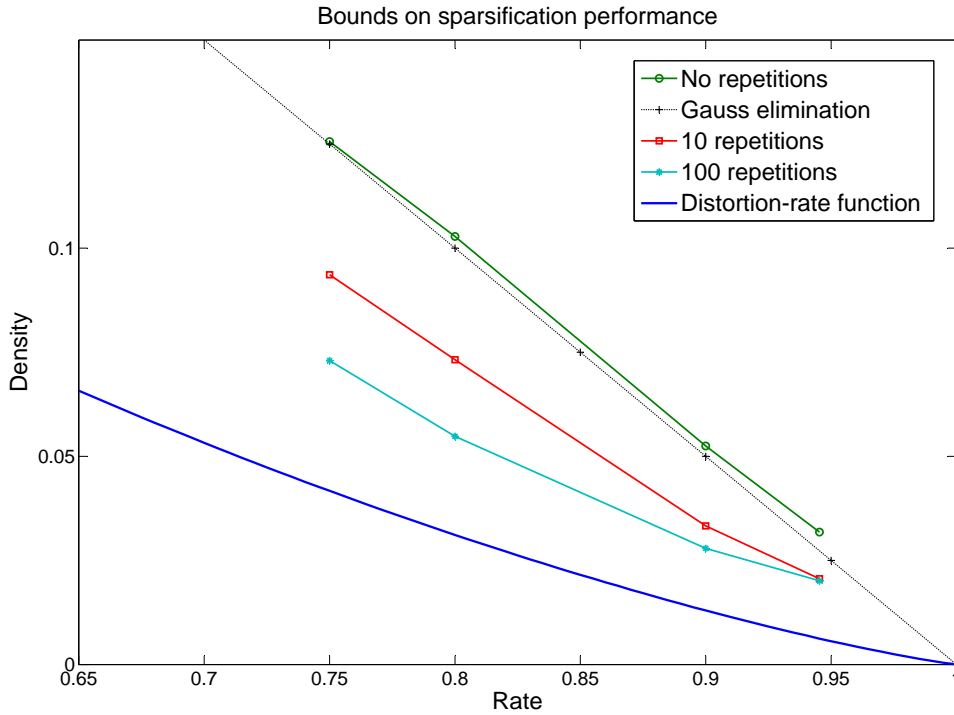


Figure 4: Approaching the distortion-rate curve using Algorithm 2.

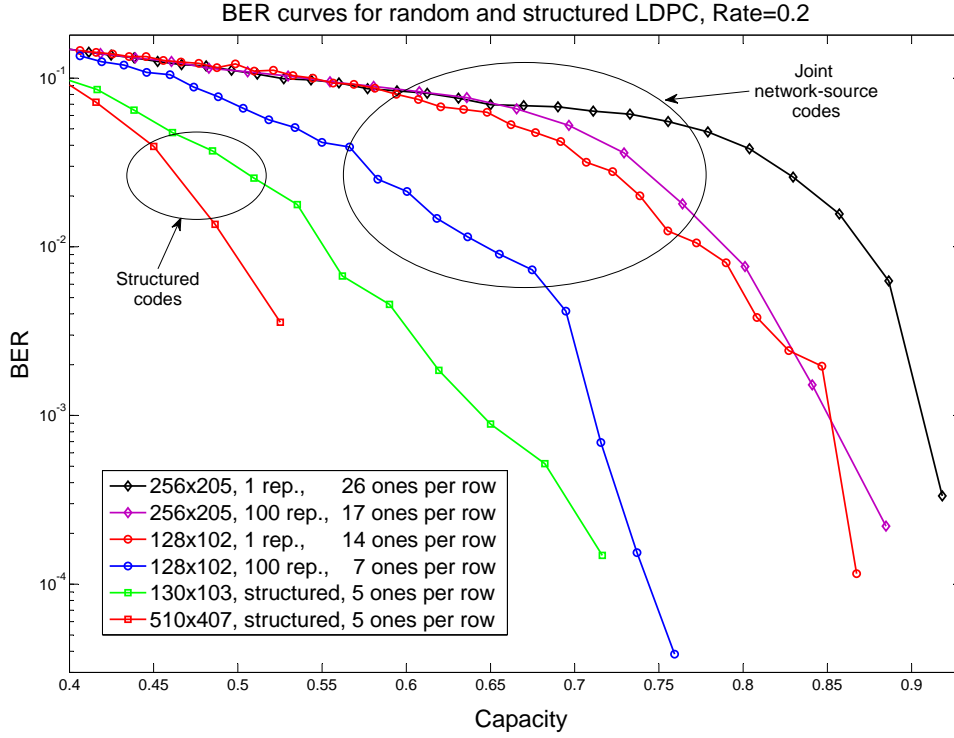


Figure 5: Bit error rates for LDPCs created by applying randomized sparsification algorithm on binary symmetric i.i.d. matrices of size $n \times nR$, $R = 0.8$ (a rate 0.2 channel code), and for LDPCs created by a structured construction.

	Graph 1 - 50 rep	Graph 2 - 50 rep	Graph 3 - 50, 100, and 200 rep
H size	22x21	23x22	21x20
H zeroes	0.87013	0.89526	0.88333
B1 size	21x20	22x19	20x18
B1 zeroes	0.0571	0.055	0.0556
H*B1 zeroes	0.0614	0.1007	0.0873
H*B1*P zeroes	0.8886	0.8741	0.8783
B2 size	21x20	22x19	20x19
B2 zeroes	0.0643	0.0574	0.0474
H*B2 zeroes	0.0636	0.0961	0.1228
H*B2*P zeroes	0.8886	0.8719	0.9098
B3 size	21x20	22x21	20x19
B3 zeroes	0.081	0.0584	0.0605
H*B3 zeroes	0.0682	0.0952	0.0977
H*B3*P zeroes	0.8841	0.913	0.9048

Figure 6: Sparsification results for deterministic coding matrices over $\text{GF}(2^4)$.

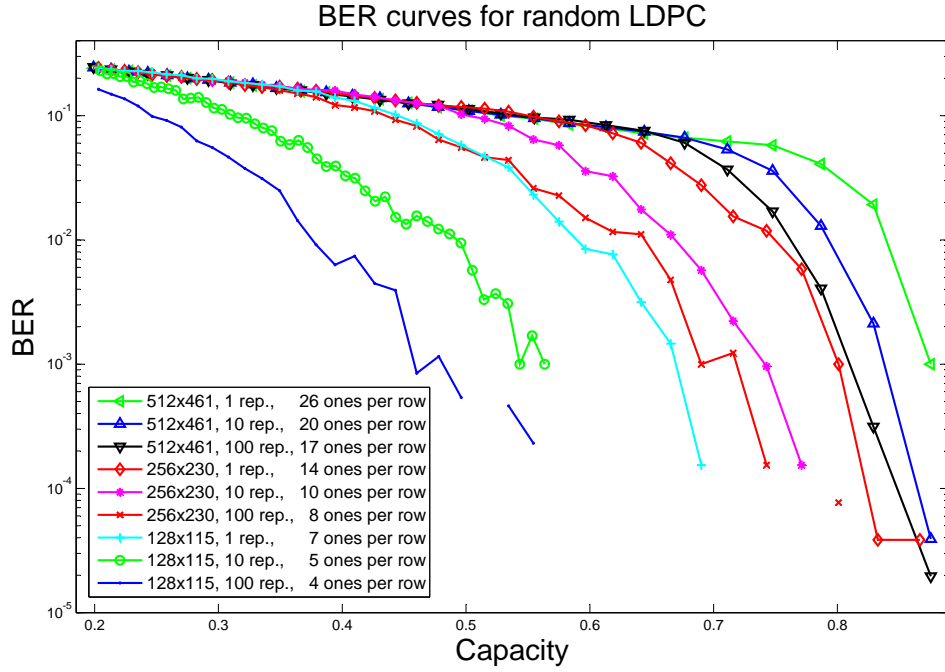


Figure 7: Bit error rates for LDPCs created by applying Algorithm 1 on binary symmetric i.i.d. matrices of size $n \times nR$, $R = 0.9$ (a rate 0.1 channel code).

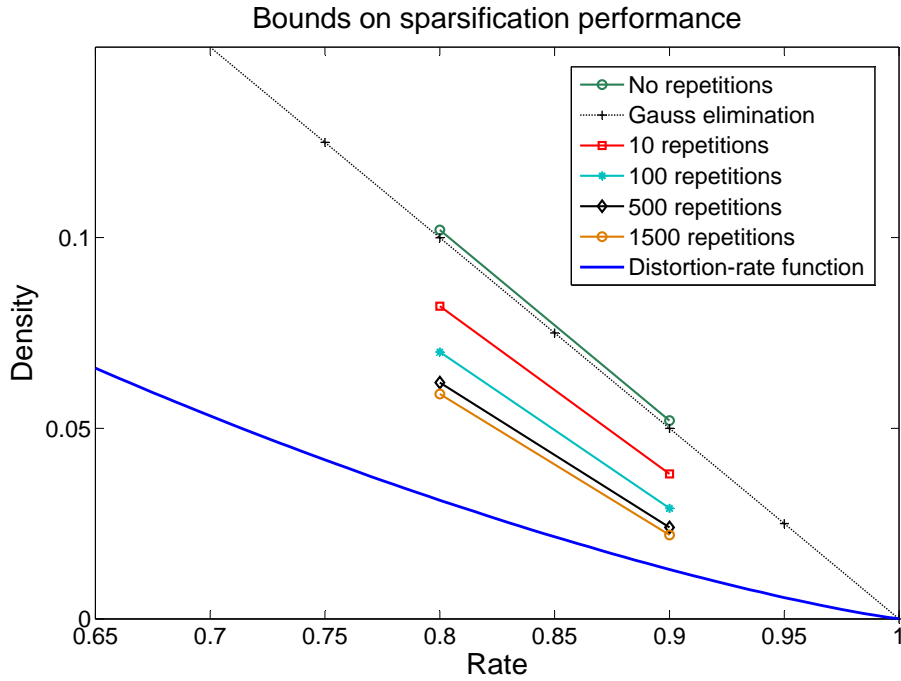


Figure 8: Approaching the minimum possible density using Algorithm 1.

A Proof of Proposition 1

Proposition 1 merely translates Corollary 1, which asserts that linear network codes such that each node t can receive $\text{maxflow}(s, t)$ independent equations on the w source symbols over some base field $GF(2^m)$, to a binary representation. Hence, each node t can receive $m \cdot \text{maxflow}(s, t)$ independent equations on the mw bits. The proof is based on the following technical claim, which shows that indeed linear equations on the elements of the source vector over $GF(2^m)$ translate to linear equations on the bits in the bit representation of the symbols. Proposition 1 will then easily follow from Corollary 1.

Claim 1. *For any set of linear equations over $F = GF(2^m)$, $\mathbf{s}M = \mathbf{y}$, where $\mathbf{s} \in F^w$, $M \in F^{w \times l}$ with rank r and $\mathbf{y} \in F^l$, there exists a set of linear equations over $GF(2)$, and a binary matrix $B \in GF(2)^{wm \times lm}$ of rank rm , such the binary representations of \mathbf{s} and \mathbf{y} , denoted \mathbf{b}_s and \mathbf{b}_y (that is, the coefficients over $GF(2)$ in the representation of the elements of F as polynomials), satisfies $\mathbf{b}_s B = \mathbf{b}_y$ over $GF(2)$.*

Proof. The system $\mathbf{s}M = \mathbf{y}$ represents r independent equations over F . Each equation is of the form

$$s_1 m_{1,i} + s_2 m_{2,i} + \dots + s_w m_{w,i} = y_i \quad (5)$$

(with a slight abuse of notation, m represents the extension degree and $m_{i,j}$ represents the i, j entry of M). Each unknown s_j can be represented as a polynomial in x ,

$$\alpha_j^0 + \alpha_j^1 x + \dots + \alpha_j^{m-1} x^{m-1},$$

where now $\{\alpha_j^l\}_{l=0}^{m-1}$ are m unknowns over $GF(2)$. The same holds for all entries of the matrix M , yet, since M is known, so do the coefficients of each $m_{i,j}$ in its polynomial representation.

Hence, (5) translates to

$$\begin{aligned}
& (\alpha_1^0 + \alpha_1^1 x + \dots + \alpha_1^{m-1} x^{m-1}) \cdot (m_{1,i}^0 + m_{1,i}^1 x + \dots + m_{1,i}^{m-1} x^{m-1}) + \dots \\
& = (y_i^0 + y_i^1 x + \dots + y_i^{m-1} x^{m-1}). \quad (6)
\end{aligned}$$

The l.h.s. of (6) has a maximal degree $2m - 2$, hence, the remainder modulo an irreducible polynomial of degree $m - 1$ is computed. This operation is linear in the coefficients of each x^i , and, as a result, the l.h.s. translates to a polynomial of degree at most $m - 1$ with coefficients which are linear functions of the unknowns $\{\alpha_j^l\}_{j=1, l=0}^{w, m-1}$. When comparing the coefficients of each power of x , the result is m linear equations for the unknowns $\{\alpha_j^l\}_{j=1, l=0}^{w, m-1}$. Clearly, r independent linear equations of the form of (5) (over $\text{GF}(2^m)$) will result in mr linear equations for $\{\alpha_j^l\}_{j=1, l=0}^{w, m-1}$ (over $\text{GF}(2)$). To see that the resulting mr equations are indeed independent, note that if $w - r$ of the unknowns s_1, \dots, s_w are given, this method must determine the values of the rest. Since the remaining r unknowns over $\text{GF}(2^m)$ are represented by mr unknowns over $\text{GF}(2)$, the mr equations over $\text{GF}(2)$ must be independent. \square

B Proof of Proposition 2

- Proof sketch.*
1. If B_t is full rank, non of its columns is the all-zeros column. Assume column j of B_t contains l ones. Then, for each i , the entry (i, j) of HB_t is the sum of l independent Bernoulli(λ/n) random variables, and the summands for the entry (i, j) are independent of those for (i', j) , for any $i \neq i'$.
 2. If B_t is full rank, non of its columns are identical. This means that the sums composing $(HB_t)(i, j)$ and $(HB_t)(i, j')$ differ in at least one random variable. Since this random variable is binary, uniform and independent of all the rest, $(HB_t)(i, j)$ and $(HB_t)(i, j')$ are also uniform and independent. For different output rows this holds trivially.
 3. The event that $(HB_t)(i, j) = 0$ is the event that an even number of the Bernoulli

random variables is 1, hence

$$\Pr((HB_t)(i, j) = 0) = \frac{1}{2} \left(1 + \left(1 - \frac{2\lambda}{n} \right)^l \right) \quad (7)$$

for all i . Assume $l = c_j n$ and compute the limit as $n \rightarrow \infty$. For large enough λ , the distribution is arbitrarily close to uniform bits.

4. We use the result of Item 3 above. Since the number of non-zero entries in column j of B_t is sub-linear in n , for large enough n it is smaller than any cn (for arbitrary small $c > 0$). Thus,

$$\frac{1}{2} \left(1 + \left(1 - \frac{2\lambda}{n} \right)^{cn} \right) \leq \Pr((HB_t)(i, j) = 0) = \frac{1}{2} \left(1 + \left(1 - \frac{2\lambda}{n} \right)^l \right) \leq 1. \quad (8)$$

Taking the limit of $n \rightarrow \infty$ we have

$$\frac{1}{2} (1 + e^{-2c\lambda}) \leq \Pr((HB_t)(i, j) = 0) \leq 1 \quad (9)$$

for arbitrarily small c , which completes the proof.

5. If $d(\mathbf{b}_j, \mathbf{b}_{j'}) = \Theta(n)$, then the difference between $(HB_t)(i, j)$ and $(HB_t)(i, j')$ is a sum of a linear number of independent Bernoulli(λ/n) random variables. This sum is a uniform bit if $\lambda = \omega(n)$, hence the random variables $(HB_t)(i, j)$ and $(HB_t)(i, j')$ are independent for each i . However, if this sum is sub-linear, then a constant λ results in an asymptotically zero probability for it to be 1 (similar to item 4 above). This means that the probability that $(HB_t)(i, j)$ and $(HB_t)(i, j')$ *differ* is small. Note that in this case, with high probability, the *sum* of $(HB_t)(i, j)$ and $(HB_t)(i, j')$ is 0, hence summing the two columns gives a sparse column.

□